Machine Learning Approaches to Understand IT Outsourcing Portfolios

Abstract

Advances in machine learning (ML) and natural language processing (NLP) holds out tremendous insights both for discovering patterns in data and for better understanding the distinct context of inter-firm arrangements. The outsourcing of information technology (IT) services poses a conundrum to the traditional theories of the firm. While there are a lot of prescriptive sourcing metrics that are geared towards the evaluation of tangible and measurable aspects of vendors and clients, much of the information that is traditionally important in making such decisions is unstructured, as it is captured through discussion and analyses summarized in contracting decisions, self-disclosed aspects of such arrangements explicated in press releases of outsourcing arrangements. NLP methods can parse through such press releases and disclosures, and learn vendor and client specific information. We train and apply our own NLP model based on deep learning methods using doc2vec, which allows users to create semi-supervised methods for representation of words. The question we ask is: *How can we employ machine learning and natural language processing to capture the difficult to elicit aspects of vendor selection in outsourcing IT services?* We find two novel constructs, *vendor-client alignment and vendor-task alignment*, that shape partner selection and the alternatives faced by clients in IT outsourcing, as opposed to agency or transaction cost considerations alone. Our method suggests that NLP and machine learning approaches provide additional insight, over and above traditionally understood variables in academic literature as well as trade and industry press, about the difficult to elicit aspects of vendor-client interaction. Implications for research and practice are discussed.

1. INTRODUCTION

Industry surveys indicate robust and continued interest in the spending on IT outsourcing. Per the recent KPMG CIO survey of 2019[1] , 41% of the 3600 respondents state they plan to increase their IT Outsourcing spending next year for the primary reason of accessing skills not available inhouse. However, a challenge in understanding outsourcing for IT services is that such outsourcing initiatives do not fit the continuum of make-vs-buy relationships prescribed in prior theories. Recent decades have likewise witnessed an unprecedented rise in vertical *de-integration* and outsourcing of complex products and services (Linder et al. 2003, Miozzo and Grimshaw 2005). A remarkable aspect of such inter-firm collaborations is that they are governed through a blend of explicit and implicit obligations enmeshed within a formal structure of exchange (Gilson et al. 2009). Such obligations can be described neither as arm's length arrangements nor relational contracting and offer a contrast to traditional theories of the firm.

Managerial and economic motivations for exchange posit that partner selection is driven by capabilities and transaction costs. Industry analysts and business press have advocated various attributes that clients may adopt in choosing vendors. It has been posited that a vendor's domain expertise, financial backing, ramp up capability and etc. should be the dominant reasons for selecting vendors. However, IT outsourcing initiatives are not standardized, and vendors are heterogeneous in their capabilities, making it difficult to generalize such advice beyond a few large and reputed IT vendors. Prior research employing a contract theoretic perspective contends that uncertainty in operations (Aubert et al. 2004) and in vendor capabilities (Banerjee and Duflo 2000), and post-contractual holdup by the vendor (e.g., Susarla et al. 2010) critically influence client's vendor selection decision. Theories of relational governance predict depth of relationships

[1] A Changing Perspective, Harvey Nash / KPMG CIO survey 2019, accessible at https://assets.kpmg/content/dam/kpmg/nl/pdf/2019/advisory/cio-survey-2019-harvey-nash-kpmg.pdf

between exchange partners that transact frequently (Susarla et al. 2020) and theories of transactional outsourcing predict that firms should diversify risks through developing a portfolio of vendors (e.g., Bakos and Brynjolfsson 1993). A networks-based perspective would, by contrast, suggest that network information be leveraged to reduce ex-ante uncertainty and ex-post hold up risk (Ravindran et al. 2015). In such a perspective the choice of a trading partner is based on not just dyadic information but a total of the information flow between the various members of the network. According to this perspective, clients would face a tradeoff between switching costs with an incumbent (see Whitten et al. 2010 for a discussion of the types of switching costs) vs. the advantages of market efficiency in adopting a new provider. Networks perspectives would suggest clients may exhibit a preference for factors such as diversity and centrality.

In the context of outsourcing of IT services, one sees an interesting pattern of both greater depth and breadth than suggested by prior literature. We see long-tail outsourcing arrangements where companies have reinvented supplier portfolio where clients prefer to deal with a smaller set of vendors (Su et al. 2015) alongside and simultaneously clients choosing to broaden their portfolio of sourcing choices through multisourcing (Su and Levina 2011) and by inviting new vendors (Cullen et al. 2005). This highlights the limits of the applicability of prior theories of sourcing, based on observed *atheoretical* sourcing arrangements. We are motivated by this empirical puzzle in understanding what are the factors that explain the seemingly atheoretical motivations for sourcing. We want to develop methods to highlight various contingencies of contracting that drive partner selection. We use cutting-edge machine learning (ML) and natural language processing (NLP) methods to discover insights about the process of client and vendor engagement in choosing exchange partners. Recent research in economics as well as management has demonstrated the novelty of insights arising from ML based approaches (Choudhury et al. 2019, Tidhar and Eisenhardt 2020). The question we ask is: *How can we employ machine learning and natural*

language processing to capture the difficult to elicit aspects of vendor selection in outsourcing IT services? A related question is whether we can consider novel sources of information that clients and vendors could have access to in navigating these complex choices? Traditionally an assessment of vendor capabilities and fit was limited to survey items and field studies (e.g., Ethiraj et al. 2005). However, the challenge is that such methods cannot be employed for archival data, nor are they scalable for a large-scale industry level analysis. Using sophisticated methods from machine learning and NLP algorithms, we build proxies to reveal the nuances of this process. We find two novel constructs, *vendor-client alignment and vendor-task alignment*, that shape partner selection and the alternatives faced by clients in IT outsourcing, as opposed to agency or transaction cost considerations alone. Thus, our approach is in line with recent research in strategy (e.g., Chowdhury et al. 2019) that has employed machine learning methods to create novel constructs that illuminate the limited applicability of prior theories and contribute to new theory building. Our work also illustrates that NLP methods can help bridge a gap into empirical inquiry into phenomenon where there was a paucity of relevant data.

The nature of ambiguity, contextual uncertainty and idiosyncrasy in IT outsourcing make it a rich field to mine for insights using unstructured data. We look at the overall market of clients in vendors engaging in complex IT services outsourcing arrangements over a 20-year period. Given our access to the comprehensive contract description and sophisticated methods and algorithms, we build proxies to reveal the nuances of this process by conceptualizing the contract announcement as a document embedding. We build upon word embeddings models that use deep learning methods in computational linguistics (Mikolov et al., 2013) to represent words as vectors. These methods can be used to represent a document, which can be any piece of text ranging from user reviews, short form content from social media, articles, books etc., and in our case the textual description about the contract from the press releases. This enables us to model the nuances of not

only how contracts are drafted depending on detailed vendor-client negotiations but also captures hard-to-elicit details about clients and vendors. For example, we can identify that a contract for hardware outsourcing may be closer in the vector space to a maintenance contract, based on details of the contracted task, as opposed to some other hardware engagements. Similarly, our method can be used to identify which vendors are closer to each other and which clients engage in structurally similar arrangements.

Since we are conceptualizing client-vendor contract descriptions as embeddings, the counterfactual we need to consider is whether the document embedding would capture the same information from mapping contracts in an inter-organizational network. Therefore, we also employ a network-based logic and consider alternate explanations from network governance. We find that our method outperforms other method of representing contracts. Another advantage of our method is that it is trained to recognize the nuances of the IT outsourcing context, since we are not using a pre-trained classifier trained over a general-purpose repository such as Wikipedia.

This study offers the following contribution: First, we focus on machine learning based logic on vendor selection in IT outsourcing which is more generalizable than a dyad specific, theory of the firm-based logic. Next, we add more nuance to inter-firm contracting literature by representing contracts as embeddings. Prior literature has looked at inter-firm networks, but such work misses the context of what transpires within the contract and focuses on the contract as a dyadic tie between symmetric network actors. Representing contracting ties as embeddings is a fairly novel approach in IS and can be broadly used in management and economics research. Finally, we calibrate the models using a sophisticated approach to predicting which vendor should be awarded a contract.

2. THEORY

Contracting Hazards in IT Outsourcing

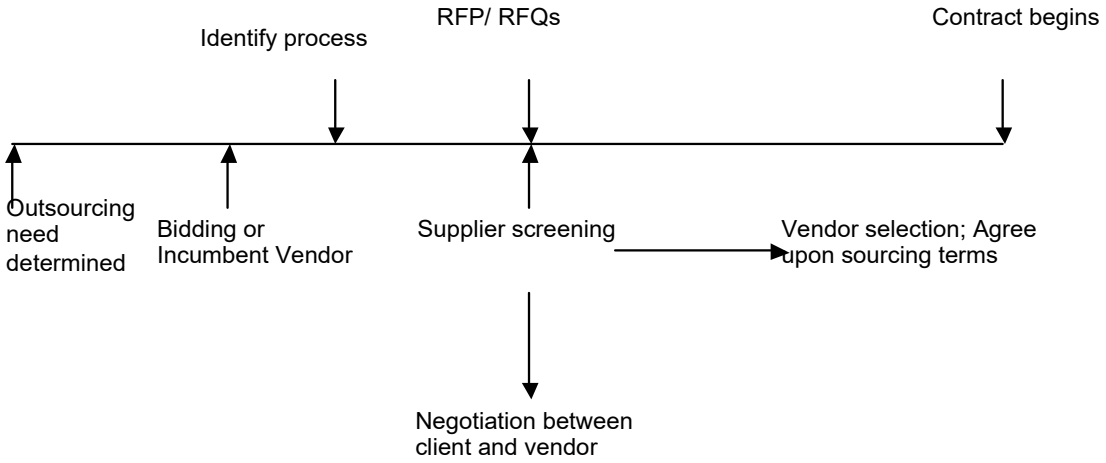
One of the fundamental characteristics in the outsourcing of IT services is that of uncertainty. The uncertainty about future technologies and prices of underlying inputs such as hardware and network infrastructure (Gurbaxani 2007), and the difficulty in defining the stream of services upfront (Banerjee and Duflo 2000) limit the ease with which parties can specify performance milestones and penalties (e.g., Whang 1992). Further, given the intangibility and firm specificity of IT services (Bresnahan, Brynjolfsson and Hitt 2002), IT outsourcing contracts, even if undertaken between the same exchange partners, need not involve the repeated exchange of identical services (Mayer and Nickerson 2005). IT outsourcing also involves considerable deployment of specific investments to tailor processes to an organizational context (Poppo and Zenger 2002). Specific investments combined with contract incompleteness make contracts for IT outsourcing fraught with ex-post opportunism (Susarla et al. 2010). For this reason, failures are common, and it can be difficult to ascribe responsibility for failure to one party alone (Banerjee and Duflo 2000). Given the complexity and substantial uncertainty characterizing IT (Barthélemy and Quélin 2006), incorporating explicit provisions to deter the threat of inefficient bargaining can make the contract too rigid to deal with future contingencies (Susarla 2012), and can lead to ex post governance costs where parties seek to appropriate quasi-rents through behavior such as mal-adaptation, haggling, set-up, and bonding costs (Susarla et al. 2009). Such behavior also increases ex post monitoring costs in ensuring that the contracting obligations are being met. Anticipating acrimonious bargaining, a vendor could also under-invest in the non-contractible, client-specific investments (Susarla 2012). At the same time, the rapidity of technological change in the IT sector implies that foregoing outsourcing deprives client companies of access to innovative capabilities (Susarla and Mukhopadhyay 2019).

Theories of Partner Selection

While substantial literature analyzes outsourcing of IT services (see Dibbern et al. 2004 and Kotlarsky et al. 2018 for comprehensive reviews), partner selection and client-vendor relationships, while being key (Liang et al. 2014 identify client partner relationships as one the main paths in the evolution of ITO literature) has received less attention. A recent paper by Wolverton et al. (2020) examines client expectations from vendors and how vendors may be compared to peers.

For large IT outsourcing contracts, partner selection is the result of an exhaustive process of negotiation. The first step is when the client company formulates a sourcing strategy. The process of supplier selection is usually a multi-stage and multi-step process, wherein vendors are invited to either submit bids or quotes in response to a request for quotation (RFQ) or request for proposal (RFQ) put out by the client company. Vendors are then selected through competitive bidding or through negotiations. Figure 1 describes the process of vendor selection.

Figure 1. Vendor Selection Process



Prior literature in outsourcing has largely examined vendor selection in outsourcing using theories of the firm such as Transaction Cost Economics (Aubert et al. 2004), Resource Based View (Espino-Rodriguez and Padron-Robaina 2006) and Property Rights Theory (Chang et al.

2017). In these theories the trading partner of choice was one that either lowered transaction costs, or offered scarce resources, or allowed control over dependent assets, respectively. What is common to these is the perspective of a firm as an independent player in a dyadic relationship.

A tradition of research in transaction cost economics posits that two fundamental issues in exchange are that of 'fundamental transformation' and 'small numbers bargaining' (Williamson 1996). When specific investments are important to an exchange, parties prefer to contract repeatedly with a smaller set of trading partners because the value of trading in bilateral exchange or with a smaller group of exchange partners outweighs the advantages of trading with others (Williamson 1996). However, in the market for IT services, the requirements and technological capabilities expected from IT are constantly evolving and at a very rapid pace. For instance, an aspect of vendor capability that is specific to IT services is that of cyber security (Dhillon et al., 2017).

Newcomers may therefore have an advantage over the incumbent who might have been slow to adopt the newer requirements. Given the changing technology and constantly decreasing input prices, clients may want to avoid lock-in into higher prices.

Network-based theories demonstrate the value of examining a firm as one that is embedded in a network of myriad partners (Beckman et al. 2004). In such a perspective the choice of a trading partner is based on not just dyadic information but a total of the information flow between the various members of the network. Client firms could utilize their existing knowledge base to facilitate information exchange, as well as improving the efficiency of future collaborations (Li and Rowley 2002, Lavie and Rosenkopf 2006). At the same time, client companies can also leverage their network positions to reach novel knowledge that otherwise does not exist in their existing partners (Rosenkopf and Nerkar 2001), as well as to diversify their portfolio to be better prepared for environmental adaptation (McGrath 2001).

There is another tradition of work using inter-firm networks to posit reputation-based explanations, rather than explanations from social capital. According to this literature, the network serves as a mechanism to transmit hard to acquire information about the quality of market participants and facilitates the rewarding of cooperative behavior (DiMaggio and Louch 1998, Raub and Weesie 1990). In IT Outsourcing, it is not uncommon to see a multi-vendor configuration where the firms aim to benefit by way of an explorative rather than exploitative nature of organizational learning (Koo et al. 2017).

Table 1: A Comparison of Different Theoretical Traditions

	Transaction Cost Economics	Relational Contracts	Market Reputation	Network Theories
Observable/ Verifiable Outcomes	Complete contract, Outcomes verifiable by third party	Incomplete contracts, outcomes observable by both parties	Contract outcomes observable by the market	Embeddedness of agents in the network is observable by others
Cost of breaching contract	Monetary damages awarded to the harmed party	Loss of future rents	Loss of reputation in the market	Loss of social capital maintained through the network
Contract Enforcement mechanism	Dispute resolution mechanisms and breach conditions specified in contract	Self-enforcement of contract	Loss of future sales	Network sanctions and stratifies agents

3. DATA AND METHODS

Measures

Our dataset is a compilation of publicly announced IT outsourcing arrangements collated by a leading industry analyst firm during the period 1989-2009, and covers a wide variety of industries, service types, regions, and sizes. The announcement typically contains a short description of the service, the duration and value and the names of the vendor and client. [2] This industry analyst has also augmented these public announcements with some of their own research to reveal other related information about the deal such as where the vendor was a subcontracted firm, whether the process

[2] Please refer to our Appendix for an example of text of such outsourcing contract announcement.

of awarding a contract was a competitive one, where the contract was signed etc. when available the analyst also tracked the current status of the contract as of 2009. There are 8072 client firms and 2569 vendor firms in our dataset with 22,039 outsourcing contracts. 32% of the observations are classified as IT outsourcing, 14% are Business Process Outsourcing, 2% Consulting, 3% Application development and maintenance and Hardware and Software Support constituted 12% of the total. 35% were classified as System Integration and the rest were the remaining 2%. To avoid service level heterogeneity, we add control variables for three service types: IT outsourcing (ITO), Business Process Outsourcing (BPO) and others. In the early parts of this time period, the outsourcing phenomenon was just picking up and as a result data is quite sparse between 1989 and 1995 (140 contracts). We therefore analyze using contracts signed after 1995. As in Ravindran et al. (2015), we exclude government contracts from our dataset as these are often awarded following standardized procurement procedures not necessarily governed by market forces. Since the choice of a prime contractor and selecting a subcontractor can be different from the decision process of a sole-sourcing contract, we eliminated multi-sourcing and subcontracting from our sample. We further divide this dataset into a calibration sample and a holdout sample: the former to train the algorithm, and the latter to predict vendor selection. The calibration sample contains contracts signed between 1996 and 2007 with 11,702 contracts, and the holdout sample contains contracts signed between 2008 and 2009 with 3,489 contracts. In one of the robustness checks, we merge this dataset with Compustat to add financial variables on size, revenue, employees and industry to the publically listed firms. These provide additional control variables that help rule out alternate causal explanations.

Table 2: Data Summary-Firm Characteristics

	Statistics	Contracts	Duration	Revenue ($million)	Employees (1000's)	EBITDA ($million)
VENDORS	Mean	6.4	4.9	9194	33.9	1537
	StdDeviation	26.5	2.6	17842	64.2	3942
	Min	1	1	0.29	0.009	-166
	Median	1	5	1410	9.5	121
	Max	312	20	89649	430	30400
	N	373	251	182	169	182
CLIENTS	Mean	1.6	6.0	16746	99.99	3289
	StdDeviation	2.1	2.8	35060	101.1	9238
	Min	1	0.4	1.8	0.03	-1808
	Median	1	5	5378	67	797
	Max	50	20	476319	443	118107

Table 3: Data Summary-Contract Characteristics

Statistics	Count	Mean	Standard Deviation	Type
Duration	1730	6.05	3.03	Continuous
Contract Value ($million)	1417	445.8	1584.9	Continuous
Annual Value ($million)	1253	63.1	192.1	Continuous
Pre-1999	2400	0.15	0.37	Binary dummy
Prior contract	2400	0.16	0.36	Binary dummy
ITO	2400	0.34	0.47	Binary dummy
BPO	2400	0.19	0.39	Binary dummy
Data Center	2400	0.05	0.21	Binary dummy
Network Mgmt	2400	0.05	0.22	Binary dummy
HW /SW Maintenance	2400	0.15	0.35	Binary dummy
System Integration	2400	0.21	0.41	Binary dummy
Others	2400	0.02	0.14	Binary dummy

Machine Learning

Increasingly, researchers in Information Systems and Management are using advanced machine learning and natural language processing (NLP) algorithms, and have demonstrate the importance of leveraging these predictive and NLP techniques (Athey 2018, Kleinberg et al 2017).

State-of-the-art machine learning algorithms have attracted significant attention lately among researchers in Management. Tidhar and Eisenhardt (2019) propose a machine learning framework to optimize company revenue models in the context of mobile apps. Choudhury et al (2021) implement machine learning algorithms to demonstrate nonlinear relationships between important variables and employee turnover at a large company, and highlight that such relationship can be easily overlooked through traditional method.

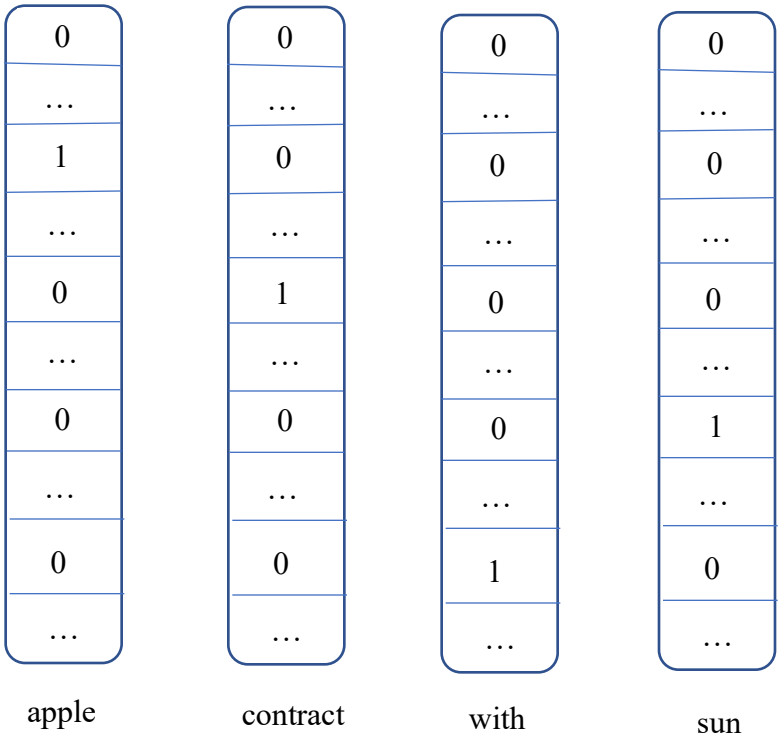
In addition, machine learning, particularly NLP, has been demonstrated to be beneficial in academic research as it facilitates exploring dimensions from "big data" that have been largely under-studies in previous literature. For example, Arts et al (2018) use text mining to construct technological similarity measure between patents, and find that knowledge spillover occurs more frequently among individuals with higher spatial proximity. Hannigan et al (2019) review recent management papers which leverage topic modeling to advance theoretical research in management. Choudhury et al (2019) leverage novel natural language processing and image recognition algorithm to examine how CEO oral communication styles influence M&A outcomes.

However, to the best of our knowledge, there is very few academic research leveraging machine learning and natural language processing in the context of outsourcing. Thus it is also our intention to fill this research gap.

Contracts as Embeddings

Most analyses employing text analytics in IS and management research have relied on the tools of sentiment analysis using a dictionary-based approach (e.g., Loughran and MacDonald 2011). Word embedding methods originate in information retrieval research and usually treat words as vector representations. Traditional machine learning methods as term frequency-inverse document frequency (TF-IDF) employ a bag of words approach that does not consider the context in which words are used. Intuitively, such natural language processing algorithms is very similar to one-hot encoding to generate a representation in which each word corresponding to a location in a vector. As illustrative purpose, consider a simple text: *Apple contracts with Sun*. This sentence corresponds to the following four vectors.

Figure 2: Representing Contracts as Embeddings



The four vectors (or *embeddings*) in figure above all have the same dimension. As an example, the first vector is for the word *apple*, thus "1" in this vector corresponds to word *apple*, and zero for all other elements of this vector. As a result, a contract description text will generate a sparse matrix. Consider an example in which there are 10,000 potential unique words. Thus for a text with 100 words, we need to generate a 10,000*100 sparse matrix.

One of the challenges of this traditional one-hot encoding is that of dimensionality as the dimension of the generated matrix is huge, making it difficult to handle. The second challenge of such one-hot encoding in traditional is how to incorporate words with similar meanings. In the context of outsourcing contracts, words such as *contract*, *agreement* or *commitment* share similar meanings. However, it is very difficult for algorithms based on one-hot encoding to identify the inherent semantic similarity among these words. Thirdly, it is also difficult for such one-hot encoding to utilize information in word order. Recent developments in natural language processing algorithms leverages neural network methods and allows us to address the dimensionality issue as

well as to identify words with similar semantic meanings. For example, skip-gram reduces the dimension of embeddings to around 300, and leverage words that are close to the focal word to understand the context in which focal word is embedded in (Mikolov et al 2013). For instance, if we were to compare the following two sentences: i) *IBM signed a contract with Oracle*, and ii) *Apple signed an agreement with Fujitsu*. By identifying that the words surrounding *contract* are similar to the words surrounding *agreement*, skip-gram can help infer that these two words "*contract"* and "*agreement"* share similar semantic meanings. As such, the generated word embeddings for "*contract"* and "*agreement*" should be relatively similar to each other.

While the above-mentioned word embedding algorithms can generate embeddings for each word, our objective in this paper is to examine and compare texts at contract level. As such, we generate embedding for each outsourcing contract text, i.e. *contract embedding* by adapting the doc2vec framework proposed by Le and Mikolov (2014). We do this by applying doc2vec on our outsourcing contract corpus to create our own contract embedding. This is because the words used in contracts for IT outsourcing can be a lot more specialized compared with the embedding developed based on a more general corpus such as Wikipedia texts. In other words, pre-trained embeddings are often based on a more general context that can be significantly different from the ones in outsourcing contract context. In this adapted doc2vec algorithm, we use 300 as the dimension of our embeddings, 5 as the window size, 15 negative samples for each good sample, initial learning rate 0.01, and 100 epochs in total. Consequently, this trained model will generate an embedding for each contract in our data (i.e. a vector of real numbers), where contracts with similar meanings will also have similar contract embeddings. For each contract, doc2vec converts it into an *embedding*. We want to note that traditional text mining techniques rely on occurrence of each unique word and cannot capture the order of words as well as the similarity between different words with similar semantic meanings. However, doc2vec represents words with vectors

with hidden information such as analogies and semantic meanings. Thus, it can better handle professional words that appear frequently in contract descriptions than text from a general background (for example Wikipedia, a commonly used training dataset for word embedding).

We then propose a similarity measure between focal outsourcing contract and previous contracts of vendor firms to conceptualize *vendor-client alignment*. Given contract embeddings generated using doc2vec, for an outsourcing contract D_m for which client firm is selecting outsourcing vendor firm, we create a cosine similarity measure to calculate the similarity between this focal outsourcing contract D_m, with D_{jn}, the n^{th} existing outsourcing contracts for a vendor firm j. If we denote this text similarity as $Simi_{im}$, we can calculate it as: $Simi_{mn} = \frac{\sum_k D_{jnk} D_{mk}}{\sqrt{\sum_k D_{jnk}^2}\sqrt{\sum_k D_{mk}^2}}$, where D_{jnk} and D_{mk} are the kth element of embedding for document D_{jn} and D_m.

We then calculate the average text similarity across all vendor firm j's outsourcing contract: $VTAlign_{jt_{ij}} = \frac{1}{N_{jt}} \sum_n Simi_{mn}$, where N_{jt} is the total number of outsourcing for vendor j before period t. This allows us to leverage vendor firm j's past outsourcing contract text to examine the propensity of being selected as vendor for contract D_m.

Lastly, we conduct text clustering for outsourcing contracts to examine the *vendor-task alignment* for each vendor. Specifically, we first employ K-means clustering algorithm to cluster outsourcing contracts based on contracts' corresponding contract embeddings based on pre-determined value of K, where contracts with similar requirement will be placed in the same cluster. We then calculate corresponding average silhouette score to measure the performance of resulting clusters. After comparing average silhouette score across different value of K, we find that K=21 gives us optimal average silhouette score. We use $VendorCluster_{jn}$ to represent the number of outsourcing contracts of vendor j in the nth cluster. Then following Blau (1977), we can calculate

the vendor-task alignment of vendor j at year t following the Herfindahl index: $VTAlign_{jt} = 1 - \sum_{n=1}^{N}\left(\frac{VendorCluster_{jn}}{\sum_w VendorCluster_{jw}}\right)^2$.

Controlling for Vendor-Client Network Measures

We construct a bipartite network, where each contract is treated as a bidirectional tie with a unit tie-strength. Since no contract exists between two clients or two vendors, the resulting network is called a bipartite or a two-mode network (Wasserman and Faust 1994) with the 8072 client firms comprising one of the modes and the 2569 vendor firms comprising the other. The network can be represented as a matrix P(I,J) where J, the number of unique vendors is 2569 and I the number of unique clients is 8072. The value of each element in matrix P, P_{ij}, represents the number of contracts between client $i \in \{1,..,I\}$ and vendor $j \in \{1,..,J\}$. A zero value indicates no contract between these two companies. The bipartite structure lets us examine the entire industry as network of vendors and clients. We measure network position of a client in terms of their stock of prior interactions in the two-mode network. Essentially, the degree centrality in the two-mode network captures the 'volume' of transactions undertaken by a firm. We construct the network structure by using a four-year moving window as the impact of outsourcing contracts is unlikely to be permanent, and the strength of the ties between client and vendor companies may diminish over time (Soda et al 2004).[3]

The degree centrality of a vendor in the network of contracts in a client's industry (and not the entire market) serves as a proxy for the vendor's experience in a particular market. In deciding whom to award an outsourcing contract to, clients are likely to prefer to deal with vendors that have abundant experience in dealing with similar clients, and vendors with skills that are similar to what is required in the current outsourcing engagement. Thus, clients perceive a vendor's depth

[3] We also used three-year, five-year and seven-year as the size of moving window. The estimation results and model performance are generally the same.

of dedicated experience positively as an indicator of the latter's deep domain knowledge and knowledge of the industry context. For clients, a potential vendor's central network position acts as guarantor that she could be expected to uphold the terms of the contractual agreement. Clients would then choose to trade with vendors with more central network positions, making the latter likely to win more contracts, further increasing their centrality in the network. The stock of prior contracts by a vendor within its industry will act as a proxy for both its reliability and ability to cater to client-specific needs. For vendors with a depth of expertise in a client's industry, earning a bad reputation due to opportunistic behavior is likely to cost much more due to information transmitted in the industry. At the same time, when a vendor has dedicated expertise in a multitude of other industries, this could also serve as an indicator of the quality and experience of the vendor.

We also calculated a vendor's diversity both in industry and service type according to the Blau (1977) index of heterogeneity, which is one minus Herfindahl index of concentration (one minus the summed square of share of contracts in each industry/service type). Diversity in network connections increases access to novel information (Burt 2004). It has been posited that different firms occupy different niches in a network whereby the niche offers access to resources and information (e.g., Podolny et al. 1996). In the market for IT outsourcing, vendors similarly occupy different niches wherein they could be generalists or specialists. Park and Podolny (2000) suggest that the highest status organizations in inter-firm networks are typically generalists that cater to a broad range of industries and diverse set of clients. For clients, diversity in a vendor's network is a good indicator that the latter have acquired multi-faced experience, both in terms of the demands of governance and in the nature of domain knowledge and application capabilities required.

When a prior relationship exists between the client and the vendor, the client has better information about the reliability and the capability of the vendor. Relational embeddedness resulting from prior ties reduces the need to include elaborate safeguards and monitoring clauses

to protect against potential opportunism (Poppo and Zenger 2002). A history of prior relationships also alleviates the fear that exchange partners will act opportunistically (Gulati 1995). Clients and vendors derive benefits in continuing to contract with each other as they can leverage the information from past relationships to counter the inherent uncertainty in outsourcing initiatives. As a measure of prior interaction, we measure the volume of past contracts between parties (Gulati 1995).

Lastly, we consider the indirect ties between client and vendor firms. In the context of bipartite outsourcing network, indirect ties are the paths with length three.[4] As posited by Gulati (1995), two companies having an indirect tie have "access to more information about each other than two firms with no such connection, and the larger the number of third partners two firms share, the more information these two firms are likely to have about each other." Ahuja (2000) argues that the network of indirect ties serves as a conduit of information, wherein each firm acts as both a recipient and a transmitter of information. In other words, a pair of vendor and client firms, as well as other firms on indirect paths between this pair of vendor and client firms, form a densely connected "local community" in which information is speedily diffused among these agents. Consequently, the vendor in such a tightly connected community is likely very concerned with their local reputation (Gulati and Gargiulo 1999). In other words, a higher number of indirect ties between a client and vendor suggest a more exposed position of the vendor and this visibility restricts bad behavior. Since firms that are concerned about future business can be trusted to uphold the terms of interaction, parties involved in a contract likely provide each other with the assurance that they can behave in a cooperative manner. As such, client firms are more likely to select vendor firms connected through more indirect ties.

Other Controls

[4] For example, client A has contract with vendor B, vendor B has contract with client C and client C has contract with vendor D. In this case, A->B->C->D is an indirect tie.

We also control for several other attributes. We employ client degree centrality as a proxy to control for client's reputation on the market. As proxies for industry similarity, we include the two-digit SIC codes of the client. We control for attributes of the service outsourced. Since IT outsourcing may have reached a level of maturity over the years, we control for contracts signed before 1999. We further incorporate client-specific, vendor specific and dyad specific unobserved random effects to control for unobserved characteristics. Furthermore, we visualize the outsourcing network in Figure 3 below, where circles represent client firms and squares represent vendors. The size of the nodes represents the degree of each firm, i.e., the larger the node is, the more outsourcing contracts this company has. The colors represent different network community structures (used only for illustration purposes and not used for estimation).

Figure 3(a). The overall network (cumulative network position as of the last contract year)

Figure 3(b). New contracts signed in 2008.

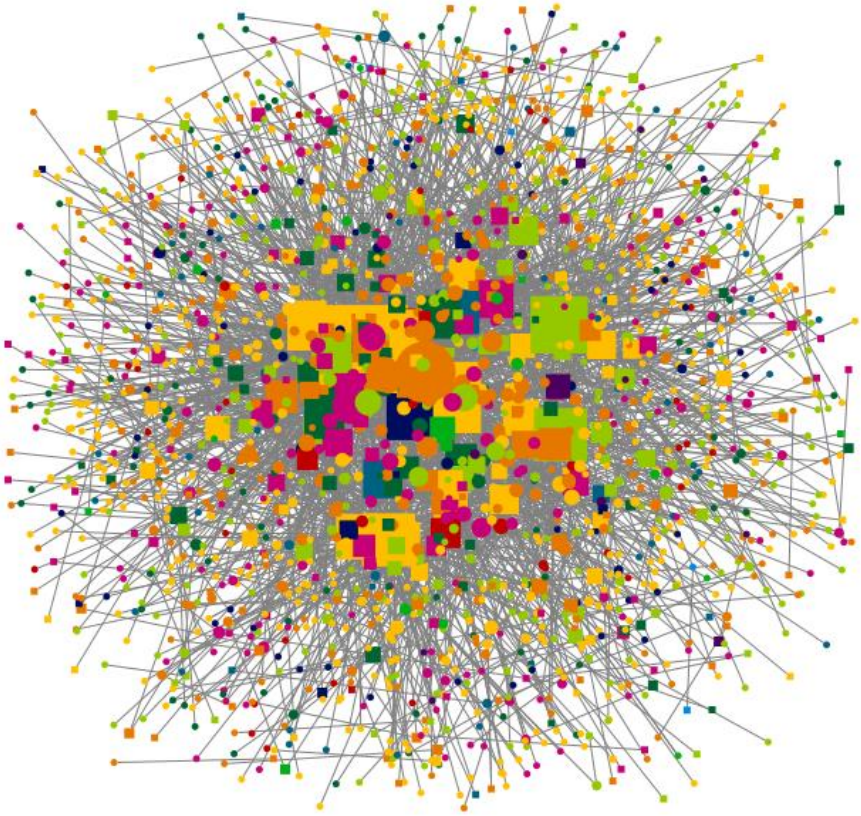
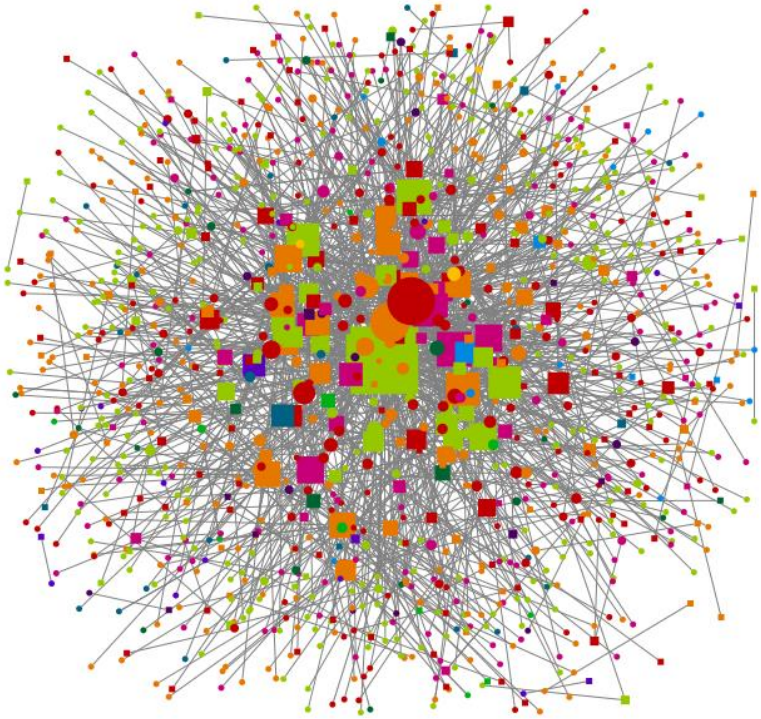


Figure 3(c). New contracts signed in 2009.



We use Table 4 below to provide summary statistics for the covariates we employ in our model.

We refer to the underlying theoretical tradition in Table 1.

Table 4. Summary of Variables Used in the Estimation

Covariates		Prior Theoretical Tradition	Descriptive Statistics	Min	Max
Vendor-specific					
$VDegIn_{jt}$	Degree of vendor j in Client's Industry in year t	Market Reputation	0.31 (4.08)	0	253
$VDegOth_{jt}$	Degree of vendor j in Other Industry in year t	Market Reputation	1.37 (14.14)	0	449
$VDivInd_{jt}$	Diversity of vendor j in terms of Industry in year t using HHI. If we use $VDivInd_{jmt}$ representing number of outsourcing contracts of vendor j in industry m till year t, this diversity is calculated as $1 - \sum_m (\frac{VDivInd_{jmt}}{\sum_p VDivInd_{jpt}})^2$	Market Reputation	0.77 (0.30)	0	0.974
$VDivServ_{jt}$	Diversity of vendor j in terms of Service Type in year t using HHI. If we use $VDivST_{jnt}$ representing number of outsourcing contracts of vendor j in service type n till year t, this diversity is calculated as $1 - \sum_n (\frac{VDivST_{jnt}}{\sum_p VDivST_{jpt}})^2$	Market Reputation	0.88 (0.20)	0	0.982
$VTAlign_{jt}$	Vendor Task Alignment of vendor j in year t, calculated using Blau index based on detected topics of contract description.	Novel theoretical construct proposed in this study	0.89 (0.03)	0	0.949
Client-specific					
$CDeg_{it}$	Degree of client i in year t	Network theories	0.54 (3.71)	0	37
Dyadic-specific					
$PInter_{ijt}$	Prior Interaction between client i and vendor j in year t, measured as number of prior contracts	Networks theories and relational contracting theories	0.003 (0.04)	0	12
$IndTie_{ijt}$	Indirect tie between client i and vendor j in year t, measured as log of number of indirect paths between i and j	Network theories	0.329(1.018)	0	11.21
$VCAlign_{ijt}$	Vendor-client alignment in year t, calculated based on similarity measure between focal contract and all previous contracts awarded to j. We then rescale it to the interval [0,1]	Novel theoretical construct proposed in this study	0.385 (0.12)	0	1

Multinomial Logistic Regression

We use a multinomial logistic regression model to examine the decision to contract with a vendor. We use i to denote client, where $i \in \{1,2,\dots,I\}$, and j to denote vendor in the market where $j \in \{1,2,\dots,J\}$. In each period t, client i decides which new vendor they will select to sign an outsourcing contract. We write the probability of a client i choosing vendor j at time t as: $\Pr(d_{ijt}=1) = \frac{\exp(\boldsymbol{V_{ijt}^s})}{\sum_{k=1}^{J}\exp(\boldsymbol{V_{ikt}^s})}$. $\boldsymbol{V_{ijt}^s}$ denotes the "business value" or the reduced transaction cost such as that of maladaptation that client firm i derives if they sign an outsourcing contract with vendor firm j at period t. A higher value of V_{ijt} indicates a higher probability of signing a contract between i and j in period t. There are three types of covariates that could influence the probability of signing a contract: client-, vendor- and dyad-specific covariates for the contract between i and j at time t represented by $\boldsymbol{X_{it}^c}$, $\boldsymbol{X_{jt}^v}$ and $\boldsymbol{X_{ijt}^d}$, respectively as presented in Table 4. Thus: $V_{ijt} = \boldsymbol{X_{it}^c\beta^c} + \boldsymbol{X_{jt}^v\beta^v} + \boldsymbol{X_{ijt}^d\beta^d}$. $\boldsymbol{\beta^c}$, $\boldsymbol{\beta^v}$ and $\boldsymbol{\beta^d}$ are the coefficients for client-specific, vendor-specific, and dyad-specific covariates respectively and c_i, v_j and d_{ij} represent client-specific, vendor specific and dyad specific unobserved characteristics respectively, we can write V_{ijt} using the function below:

$$V_{ijt}^S = (\boldsymbol{X_{it}^c\beta^c} + c_i^s) + (\boldsymbol{X_{jt}^v\beta^v} + v_j^s) + (\boldsymbol{X_{ijt}^d\beta^d} + d_{ij}^s)$$

We assume that these random effects are independent and follow normal distribution: $c_i \sim N(0,\sigma_c^2)$, $v_j \sim N(0,\sigma_v^2)$, and $d_{ij} \sim N(0,\sigma_d^2)$. Thus $X_{it}^c\beta^c + c_i$ is the client-specific effect, $X_{jt}^v\beta^v + v_j$ is the vendor-specific effect and $X_{ijt}^d\beta^d + d_{ij}$ is the dyad-specific effect. Note that these unobserved variables can be identified because we observe multiple contracts from the same vendor j and multiple contracts from the same client i. We include a control variable for

prior relationship[5] between focal vendor and client firms as suggested in the literature. Note that for the observations in which client i did not sign any outsourcing contract in year t, we use information of next outsourcing contract client i signed in subsequent years to construct vendor-client alignment. Thus, the likelihood function we estimate in this model is as equation below, where $D_{ijt} = 1$ if i signed a contract with j at period t in our dataset:

$$logL = \sum_{ijt} D_{ijt} \log\left(\frac{\exp\left((\boldsymbol{X_{it}^{c}\beta^{c}} + c_i) + (\boldsymbol{X_{jt}^{v}\beta^{v}} + v_j) + (\boldsymbol{X_{ijt}^{d}\beta^{d}} + d_{ij})\right)}{1 + \sum_{k=1}^{J} \exp\left((\boldsymbol{X_{it}^{c}\beta^{c}} + c_i) + (\boldsymbol{X_{kt}^{v}\beta^{v}} + v_j) + (\boldsymbol{X_{ikt}^{d}\beta^{d}} + d_{ik})\right)}\right)$$

4. RESULTS

We first present the estimation results below.

Table 5. Estimation Results.

[5] The *prior interaction* here is calculated as the number of prior contracts between client and vendor, irrespective of the contract type.

	Coefficient
Vendor-specific Covariates	
Vendor Deg. Centrality at t-1	0.019***
Vendor Diversity (Industry) at t-1	0.051**
Vendor Diversity (Service Type) at t-1	0.023*
Vendor-Task Alignment	0.094**
Dyadic-specific Covariates	
Vendor-Client Alignment	0.172***
Prior Interaction	2.693**
Prior Interaction*Vendor Degree	-0.007
Prior Interaction*Client Degree	-0.022*
Indirect Tie	0.095***
Indirect Tie *Vendor Degree	0.000
Indirect Tie *Client Degree	-0.001***
Client-specific and Control Covariates	
Client Degree at t-1	0.003
Pre 99	-0.176***
2-digit SIC codes	included
ITO	0.219***
BPO	0.105***
#Observations	72,012

***, ** and * denote that the 99% credible interval, the 95% credible interval, and the 90% credible interval, respectively, does not include zero. The benchmark service type case is *Other*. 2-digit industry control variables have not been reported in the interest of space.

Estimation results suggests that both NLP derived variables, vendor-task alignment, and vendor-client alignment, have significant impact. For vendor-task alignment, a greater level increases vendor's chance of obtaining a contract ($\beta = 0.094$, p-value<0.05). Specifically, when vendor-task alignment increases by 1%, vendor's chance of being awarded outsourcing contract increases by 8.34%. In addition, the coefficients of client-vendor alignment are positive and significant ($\beta = 0.172$, p-value<0.01). This suggests that vendor firms who are more suitable for this outsourcing contract is more likely to be awarded the outsourcing contract. When vendor-client alignment increases by 1%, vendor's chance of being awarded outsourcing contract increases by 18.5%. This indicates that when client firms are looking for a new vendor, selecting

a vendor firm with better fit significantly reduces risk. In other words, when this vendor has a competitive advantage on the focal outsourcing contract compared with other vendors on the market, this firm is more likely to be awarded the outsourcing contract.

Model Performance

In this section, we compare the predictive power of models with and without the NLP based variables in next section. As we will demonstrate below, NLP variables can significantly enhance the predictive power of models. In the last three decades, outsourcing has become the biggest channel to source scarce capabilities and has formed a significant activity of the CIO. Contracts for outsourcing have also become much more complicated over time with the range and complexity of services being outsourced. The business press has suggested various vendor selection criteria specific to service or software (for example Gartner's Vendor evaluation model for ERP, 2010[6]) including functional fit, culture etc. Machine learning based methods could elicit difficult to measure aspects such as complementarity between outsourcing and investments in internal IT (Han and Mithas 2013) or organizational learning in inter-firm arrangements (e.g., Dekker and Abbeele. 2010). Our method suggests that machine learning approaches provide additional insight, over and above traditionally understood variables in academic literature as well as trade and industry press, about the difficult to elicit aspects of vendor-client interaction.

Cross Validations

To better examine the performance of our model, we use two variables as the objectives of this test. We first consider the accuracy of predicting vendor signing a contract in a specific year (a binary variable in which the value is one if focal client signed a contract in a specific year and zero otherwise).

Table 6. Predicting whether client signed a contract in a specific year

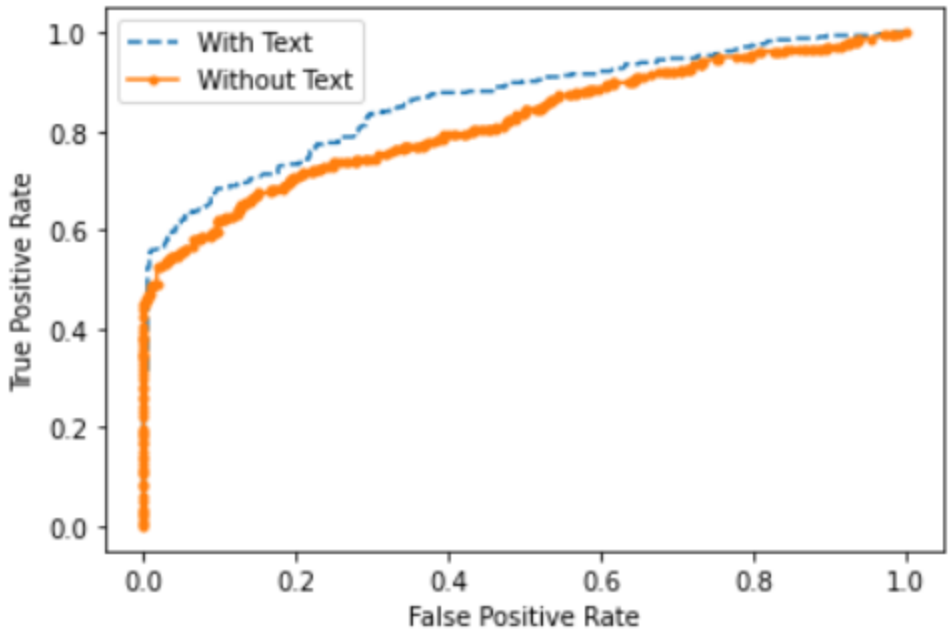
[6] Use a Vendor Evaluation Model to Select ERP Vendors and Software Denise Ganly, Michael Dunne, Mike Blechar https://www.gartner.com/en/documents/1435426/use-a-vendor-evaluation-model-to-select-erp-vendors-and-

	With both text-based variable	With only Vendor-Client Alignment	With only Vendor-Task Alignment	With no text-based variable
AUC of ROC	0.868	0.861	0.822	0.819
F1-score	0.744	0.731	0.730	0.728
precision	0.836	0.835	0.829	0.837
recall	0.779	0.782	0.743	0.728

Table 6 above compares the model performance of predicting whether a client signed a contract in a specific year. We consider four models: 1) not include the two text-based covariates; 2) only include Vendor-Client Alignment; 3) only include Vendor-task Alignment; and 4) include both text-based variables. This table demonstrates the importance of including contract text description when predicting whether a client signed a contract in a specific year. As we can see from table above, including text-based variables generally help improve the prediction of whether a client signed a contract or not. Including the Vendor-Client alignment variable has a much stronger impact compared with Vendor-task alignment variable.

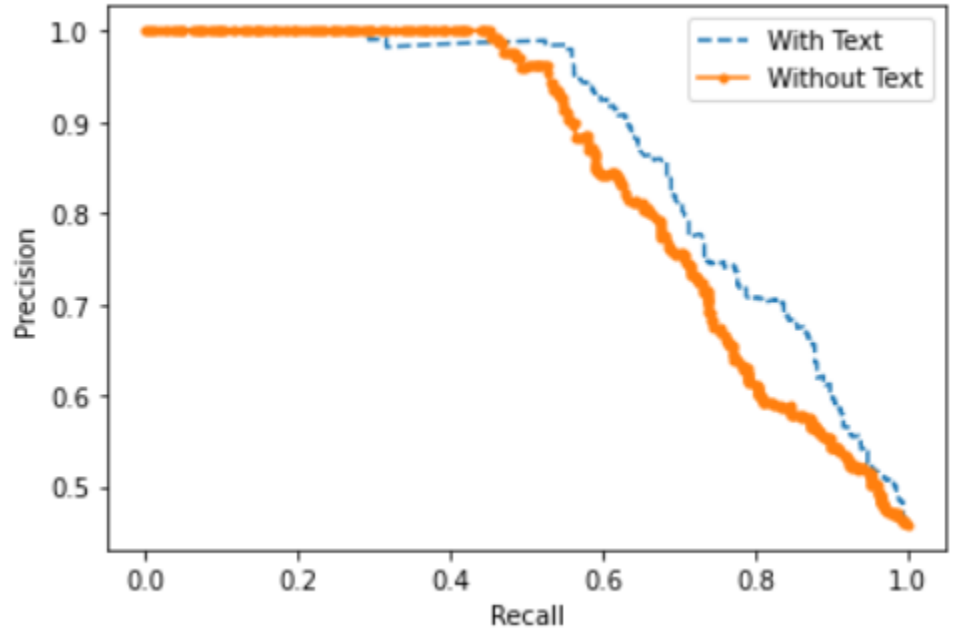
We further employ ROC and precision-recall curve to visually demonstrate the model improvement with text-based variables. For the figures below, we only compare our full model with the model without text-based variables for simplicity, as it is difficult to distinguish multiple lines in the same figure.

Figure 4. Comparing model performance using ROC for Prob of Signing Contract



As the observation on client signing contract is not perfectly balanced (roughly 75% of the time clients do not sign a contract in a specific year), it is also useful to check the performance of our model using precision-recall curve.

Figure 5. Comparing model performance using precision-recall curve



Next, we examine the accuracy of predicting the correct vendor conditional on client signed a contract in this year. Specifically, we can use the multinomial logistic model to

calculate the probability of a vendor being selected by focal client. We then can rank the vendors based on this calculated probability. Then we calculate the percentage that the client's chosen vendor in data is the top 1, top 2 and top 3 most probable vendors based on our model prediction. We also calculate the cross-entropy of each model in table below. Here cross-entropy is calculated as $\sum_{j=1}^{n} d_j \log(p_j)$, where $d_j = 1$ if vendor j is chosen by focal client in data. p_j is the model predicted probability of choosing vendor j. We demonstrate our results in table below.

Table 7. Model Comparison for Which Vendor to Choose

	With both text-based variable	**With only Vendor-Client Alignment**	**With only Vendor-task Alignment**	**With no text-based variable**
Top 1	0.386	0.386	0.351	0.351
Top 2	0.574	0.571	0.546	0.545
Top 3	0.674	0.674	0.677	0.663
Cross entropy	1032.57	1036.72	1097.05	1098.00

As we can see from table above, for model with both text-based variables, the cross entropy is the lowest. This model also has the highest accuracy rate in terms of correctly predicting selected vendor in terms of top 1 or top 2 choices. This again demonstrates the importance of including text-based variables in our model.

Next, we demonstrate the improvement of our proposed network embedding compared with a more general word embedding using Wikipedia texts.[7] To generate contract embedding using pre-trained word embedding from Wikipedia, we take the average of embeddings of all words in each document. Like previous performance comparison tests, we first compare the predictive power of whether a client signed a contract in a specific year:

Table 8. Predicting whether client signed a contract in a specific year.

	Full model	**Wikipedia Embedding**
AUC of ROC	0.868	0.857

[7] https://fasttext.cc/docs/en/pretrained-vectors.html

F1-score	0.744	0.753
precision	0.836	0.829
recall	0.779	0.751

As we can see from table above, our word embedding outperform the one by Wikipedia regarding ROC, precision, and recall. We then move on to compare our word embedding with embedding based on Wikipedia texts on the prediction of which vendor to choose conditional on client will select a vendor in this specific year.

Table 9. Model Comparison for which vendor to choose.

	Full model	**Wikipedia Embedding**
Top 1	0.386	0.381
Top 2	0.574	0.568
Top 3	0.674	0.678
Cross entropy	1032.57	1037.70

As we can see, our word embedding outperforms the one using Wikipedia embedding in terms of cross entropy, Top 1 and Top 2 most probable predicted vendor.

Comparisons with other methods

In the table below, we further compare model performance of the multinomial logistic regression we presented before with other state-of-the-art machine learning algorithms. Specifically, we evaluate the performance of random forest, XGBoost and Neural Network in table below. For neural network, we use three hidden layers, and there are 50 neurons in each layer. We set learning rate to 0.1 with 200 epochs. We do want to note that compared with these alternative machine learning algorithms employed in this table, multinomial logistic model we employed in the main analysis is much superior in terms of interpretability as we can directly examine the impact of different features on vendor selection decision through the coefficients. Notice that we include all features in these algorithms (both network-based and NLP based covariates).

Table 10.1. Model Performance Comparison with Other Algorithms—Predicting Whether Client Signed a Contract in a Specific Year

	Full Model	**Random Forest**	**XGBoost**	**Neural Network**

AUC of ROC	0.868	0.861	0.853	0.829
F1-score	0.744	0.744	0.743	0.715
precision	0.836	0.816	0.821	0.783
recall	0.779	0.765	0.768	0.719

Table 10.2. Model Performance Comparison with Other Algorithms—Predicting Which Vendor to Choose

	Full model	**Random Forest**	**XGBoost**	**Neural Network**
Top 1	0.386	0.377	0.369	0.341
Top 2	0.574	0.576	0.540	0.528
Top 3	0.674	0.683	0.649	0.657
Cross entropy	1032.57	1081.37	937.47	1094.11

As we can see from table above, the multinomial logistic model we examine as our main model generally has similar or superior predictive power in comparison with other advanced machine learning algorithms. However, multinomial logistic model is superior in terms of interpretability.

5. DISCUSSION

Implications for Research

In the last three decades, outsourcing has become the biggest channel to source scarce capabilities across firm boundaries as well as a source of innovative potential. While traditional theories of the firm would suggest that firms employ transactional sourcing arrangements, industry evidence suggests that contracts for outsourcing have become considerably complicated over time (Deloitte Insights, 2012) with a range and complexity of services being outsourced. While the business press has suggested various vendor selection criteria specific to service or software (for e.g., Gartner's Vendor evaluation model for ERP, 2010[8]) including functional fit, culture etc., such advice only applies to a very small fraction of vendors and does not address the prevalence of sourcing for

[8] Use a Vendor Evaluation Model to Select ERP Vendors and Software Denise Ganly, Michael Dunne, Mike Blechar, 13, Sept. 2010 accessible at https://www.gartner.com/en/documents/1435426/use-a-vendor-evaluation-model-to-select-erp-vendors-and-

innovation (Gilson et al. 2009). Substantial evidence points to an empirical puzzle in understanding IT outsourcing in that the overall pattern of contracting exhibits both features of transactional sourcing as well sourcing for innovation (Su et al. 2016). It is equally puzzling whether the choice of outsourcing contracts and vendors is driven by broadly theoretical factors or bandwagon style effects driven by pronouncements from outsourcing advisory firms and consultants. Given there are broad factors that govern the choice of a vendor, we seek an answer to the empirical puzzle of "how do client firms utilize the plethora of theoretical and industry-driven considerations in selecting vendors?" In seeking answers to this question, we also need to grapple with what are the various sources of empirical proxies that shed insight into this phenomenon. An ignored yet valuable and generalizable source of information in vendor selection is to analyze unstructured language from contract filings to provide more insight into the vendor and client matching process.

In this paper, we model a client firm's vendor selection choice using NLP and machine learning methods that incorporate nuances of contracts and market actors by viewing them as embeddings. Our results are relevant in a plethora of interfirm settings such as alliances, board interlocks, open innovation etc. While it is possible to employ content analyses of textual descriptions of IT outsourcing announcements and press releases, conduct detailed ethnographic studies of client and vendor interactions, or alternatively surveys of contract choice considerations from vendors and clients, such methods cannot be scalable to an entire market and a 20-year history of complex IT outsourcing deals. Traditionally, an assessment of vendor capabilities and fit was limited to survey items and field studies (e.g., Ethiraj et al. 2005). However, the challenge is that such methods cannot be employed for archival data, nor are they scalable for a large-scale industry level analysis.

We find two new constructs (i.e., vendor-client alignment and vendor-task alignment) to elucidate the nuances of matching both between clients and vendors and between vendors and

tasks. A combination of NLP methods combined with survey or field studies could further provide insight into the nuances of IT outsourcing arrangements such as complementarity between outsourcing and investments in internal IT (Han and Mithas 2013) or organizational learning in inter-firm arrangements (e.g., Dekker and Abbeele 2010). Our method suggests that machine learning approaches provide additional insight, over and above traditionally understood variables in academic literature as well as trade and industry press, about the difficult to elicit aspects of vendor-client interaction. Most recent work employing NLP methods on unstructured data in strategy and management has employed methods such as topic models (Chowdhury et al. 2019), while our approach, building on embedding methods, which is more sensitive to the context of a word, unlike topic models employing Latent Dirichlet Allocation (LDA) that consider only co-occurences of words. Another contribution of our work is to posit an alternative to literature on inter-firm arrangements that has mostly considered either network based logics in guiding partner selection (e.g., Broschak and Block 2013) or transaction cost or RBV considerations alone.

Implications for Practice

One of the primary implications of this work is that visualizing contracts and task descriptions as embeddings could provide a means to understand competitive differentiation or segmentation. Given the widespread prevalence of algorithmic decision making, it needs to be recognized that decisions about who to contract with and what services to outsource are themselves influenced by machine learning and AI based decision making. In the area of corporate disclosure, it has been posited that firms are preparing filings that are more amenable to machine parsing and processing (Cao et al. 2020). Likewise, machine translation on digital platforms has been shown to substantially increase international trade (Brynjolfsson et al. 2019). Service provider firms may seek to actively differentiate themselves by highlighting selected language in an attempt that they be assigned the right labels or niches by other market participants, which affects the aggregate

structure of competition in the IT services market. When algorithmic screening pushes vendors to specialize in different market segments, such differentiation will impact the trajectories of the emergence of clusters in the market for IT services. Another implication from this study is the role of machine learning and NLP based approaches in determining growth strategies and in revenue generation efforts for vendors. Given the need for substantial effort needed for acquiring and maintaining exchange relationships, it will be difficult for a vendor to enter into a variety of contracting arrangements if it has been unable to fulfill prior contractual obligations.

Limitations

Our work has some limitations. Our dataset consists of publicly announced outsourcing contracts alone. However, because the contracts that are not publicly announced are unlikely to be known by clients, we believe clients mainly collect information through these publicly announced contracts. Due to data limitations, we cannot distinguish between the process whereby a client selects a set of preferred vendors from a client's final decision process. Thus, the estimates reflect the combined effects from these two processes. Distinguishing the impact of covariates on these two processes also goes beyond the scope of this paper, thus we leave this for future research. We also acknowledge the data limitation that we do not observe the outsourcing contract outcome in this research, i.e., whether the outsourcing contract failed. Unfortunately, this information is not publicly available in most cases.

6. CONCLUSIONS

While prior research has studied the impact of partner selection, very limited work exists on incorporating natural language processing into an inquiry of inter-firm contracting arrangements. This question is particularly salient in the market for IT outsourcing, which despite robust growth, is characterized by a high degree of ex ante uncertainty and potential for holdup, wherein contract

cancellations and failures are common. Our research suggests several directions for the future. Future work could investigate the dynamics of business and task requirements from IT vendors have changed over the past decade with AI innovation and platform economics in the IT industry. Since data could be sparse on some aspects of inter-firm arrangements, another extension will be to consider low data machine learning methods to learn more efficiently from training data.

References

Anderson, SW, HC Dekker. 2005. Management control for market transactions: The relation between transaction characteristics, incomplete contract design, and subsequent performance. *Management Science*, 51(12): 1734-52

Argyres, N. S., J. P. Liebeskind. 1999. Contractual Commitments, Bargaining Power, and Governance Inseparability: Incorporating History into Transaction Cost Theory. *Academy of management review,* 24(1): 49-63.

Arora, A., J. Asundi. 1999. Quality Certification and the Economics of Contract Software Development A Study of the Indian Software Industry. *National Bureau of Economic Research*, No. w7260.

Arts, S., Cassiman, B., Gomez, J. C. 2018. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.

Athey, S. 2018. The impact of machine learning on economics. In A. K. Agrawal, J. Gans, A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (in press). Chicago, IL: University of Chicago Press.

Aubert, B. A., Rivard, S., & Patry, M. (2004). A transaction cost model of IT outsourcing. *Information & Management*, 41(7), 921-932.

Bakos, J. Y., E. Brynjolfsson. 1993. Information Technology, Incentives, and the Optimal Number of Suppliers. *Journal of Management Information Systems*, *10*(2): 37–53.

Banerjee, A. V., Duflo, E. 2000. Reputation Effects and the Limits of Contracting: A Study of the Indian Software Industry. *Quarterly Journal of Economics*, 115(3): 989-1017.

Blau, P. M. 1977. Inequality and Heterogeneity: A Primitive Theory of Social Structure (Vol. 7). New York: Free Press.

Bojanowski, E. Grave, A. Joulin, T. Mikolov, 2017. Enriching Word Vectors with Subword Information, *Transactions of the Association of Computational Linguistics*

Braun, M., A. Bonfrer. 2011.Scalable Inference of Customer Similarities from Interactions Data using Dirichlet Processes. *Marketing Science* 30(3): 513–531

Brynjolfsson, E., X. Hui, M. Liu. 2019. Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform. *Management Science*. 65(12): 5449–5460

Choudhury, P., R. T. Allen, M. G. Endres, 2021. Machine learning for pattern discovery in management research. *Strategic Management Journal*. 42(1): 32-57

Choudhury, P., D. Wang, N. A. Carlson, T. Khanna. 2019. Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, 40(11): 1705-1732

Cullen, Sedden and L.P. Willcocks, 2005. IT outsourcing configuration: Research into defining and designing outsourcing arrangements. *Journal of Strategic Information Systems*. Volume 14, Issue 4, December 2005, Pages 357-387

Deloitte Insights. 2012. Why IT Outsourcing May Be Riskier Than Ever. *Wall Street Journal*. Retrieved online from http://deloitte.wsj.com/cio/2012/07/09/why-it-outsourcing-may-be-riskier-than-ever/

Dekker H.C., A. Abbeele. 2010. Organizational Learning and Interfirm Control: The Effects of Partner Search and Prior Exchange Experiences. *Organization Science*, 21(6): 1233-1250

Dhillon, G., Syed, R., & de Sá-Soares, F. (2017). Information security concerns in IT outsourcing: Identifying (in) congruence between clients and vendors. *Information & Management*, 54(4), 452-464.

Ethiraj, S. K., P. Kale, M. S. Krishnan, and J. V. Singh. 2005. Where do capabilities come from and how do they matter? A study in the software services industry. *Strategic Management Journal,* 26 (1): 25-45.

Fader, P.S., B.G.S. Hardie. 1996. Modeling Consumer Choice among SKUs. *Journal of Marketing Research*, 33(4): 442-452.

Forrester. 2011. The Forrester Wave™: Global IT Infrastructure Outsourcing, Q1 2011, *Forrester Research*.

Ghose, A., P. G. Ipeirotis, and B. Li. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, *31*(3): 493-520.

Gilson, R., C. Sabel and R. Scott. 2009. Contracting for Innovation: Vertical Disintegration and Interfirm Collaboration. *Columbia Law Review*, 109: 431-502

Gopal, A., K. Sivaramakrishnan, M. S. Krishnan and T. Mukhopadhyay. 2003. Contracts in Offshore Software Development: An Empirical Analysis. *Management Science,* 49(12): 1671-1683.

Gulati, R. 1995. Does Familiarity Breed Trust? The Implications of Repeated Ties on Contractual Choice in Alliances. *Academy of Management Journal*, 38: 85-112

Han, K., S. Mithas. 2013. Information Technology Outsourcing and Non-IT Operating Costs: An Empirical Investigation. *MIS Quarterly*. 37(1): 315-331.

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Jennings, P. D. 2019. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, 13(2), 586-632.

Hansen, M. T. 1999. The Search-Transfer Problem: The role of Weak Ties in Sharing Knowledge Across Organization Subunits. *Administrative science quarterly*, *44*(1): 82-111

Hunter, D., S. Goodreau and M. Handcock. 2008. Goodness of Fit of Social Network Models. *Journal of American Statistics Association,* 103(481): 248-258.

Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, 2016. FastText.zip: Compressing text classification models

Joulin, E. Grave, P. Bojanowski, T. Mikolov. 2016. Bag of Tricks for Efficient Text Classification, in *Proceedings of the 15th Conference of the European Chapter of the Association of Computational Linguistics*

King, G., L. Zeng. 2001. Logistic Regression in Rare Events Data. *Political analysis*, *9*(2), 137-163.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. 2017. Human decisions and machine predictions. *Quarterly Journal of Economics, 133*(1), 237-293.

Koo, Y., Lee, J. N., Heng, C. S., & Park, J. (2017). Effect of multi-vendor outsourcing on organizational learning: A social relation perspective. *Information & Management*, 54(3), 396-413.

Lacity, M., L. Willcocks. 2013. Outsourcing Business Processes for Innovation. *MIT Sloan Management Review*. 54(3): 63-69

Le, Q., Mikolov, T. 2014. Distributed Representations of Sentences and Documents. *International conference on machine learning.* Pp. 1188-1196.

Liang, H., Wang, J. J., Xue, Y., & Cui, X. (2016). IT outsourcing research from 1992 to 2013: A literature review based on main path analysis. *Information & Management*, 53(2), 227-251.

Levina N., J. Ross. 2003. From the Vendor's Perspective: Exploring the Value Proposition in Information Technology Outsourcing. *MIS Quarterly*. 27(3): 331-364

Linder, J.C., S. Jarvenpaa, T.H. Davenport. 2003. Toward an Innovation Sourcing Strategy. *MIT Sloan Management Review*, Summer, 43–49

Lu, Y., K. Jerath and P. Singh. 2013. Emergence of Opinion Leaders in a Networked Online Community. *Management Science,* 2013, 59(8): 1783-1799.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miozzo, M., D. Grimshaw. 2005. Modularity and Innovation in Knowledge-Intensive Business Services: IT outsourcing in Germany and the UK. *Research Policy* 34(9): 1419-1439

Oh, W, M.J. Gallivan, J. Kim. 2014. The Market's Perception of the Transactional Risks of Information Technology Outsourcing Announcements. *Journal of Man. Inf. Sys*. 22(4): 271-303

Poppo, L., T. Zenger. 1998. Testing Alternative Theories of the Firm: Transaction Cost, Knowledge-based, and Measurement Explanations for Make-or-buy Decisions in Information Services. *Strategic management journal*, *19*(9): 853-877.

Poppo, L., T. Zenger. 2002. Do Formal Contracts and Relational Governance function as substitutes or complements? *Strategic Management Journal*, 23(8): 707-725.

Ravindran, K., A. Susarla, D. Mani, V. Gurbaxani. 2015. Social Capital and Contract Duration in Buyer-Supplier Networks for Information Technology Outsourcing. *Information Systems Research*. 26(2): 379-397

Ruckman, K., N. Saraf, V. Sambamurthy. 2015. Market Positioning by IT Service Vendors Through Imitation. *Information Systems Research* 26(1):100-126

Shmueli, G., O. R. Koppius. 2011. Predictive Analytics in Information Systems Research. *MIS Quarterly*. 35(3):553-572.

Su, Ning; Levina, Natalia; Ross, Jeanne W. 2016. The Long-Tail Strategy of IT Outsourcing. *MIT Sloan Management Review*; Cambridge Vol. 57, Iss. 2, (Winter 2016): 81-89.

N. Su and N. Levina. 2011. Global Multisourcing Strategy: Integrating Learning From Manufacturing Into IT Service Outsourcing. *IEEE Transactions on Engineering Management* 58, no. 4 (November 2011): 717-729.

Susarla, A., R. Subramanyam, P. Karhade. 2010. Contractual Provisions to Mitigate Holdup: Evidence from Information Technology Outsourcing. *Information Systems Research,* 21(1): 37-55

Susarla A, M. Holzhacker, R. Krishnan. 2020. Calculative Trust and Interfirm Contracts. Forthcoming, *Management Science*

Tidhar, R., Eisenhardt, K. M. 2020. Get Rich or Die Trying. Finding Revenue Model Fit using Machine Learning and Multiple Cases. *Strategic Management Journal*, 41(7), 1245-1273.

Uzzi, B. 1997. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. *Administrative Science Quarterly*, 42(1), 35-67.

Whitten, D., Chakrabarty, S., & Wakefield, R. (2010). The strategic choice to continue outsourcing, switch vendors, or backsource: Do switching costs matter?. *Information & Management*, 47(3), 167-175.

Wolverton, C. C., Hirschheim, R., Black, W. C., & Burleson, J. (2020). Outsourcing success in the eye of the beholder: Examining the impact of expectation confirmation theory on IT outsourcing. *Information & Management*, 57(6), 103236..

APPENDIX

APPENDIX I. MCMC Inference for Hierarchical Bayesian Model

In this Appendix, we describe estimation details of the multinomial logistic regression with unobserved heterogeneity using MCMC approach. As MCMC approach involves iteratively updating values of parameters, we use superscript n to represent the parameter values in the next iteration.

Step 1: Generating $\boldsymbol{\beta}^n$, the coefficients in multinomial logistic model that is homogenous across firms.

$$\boldsymbol{\beta}^n|\boldsymbol{\beta}, c_i\ , v_j\ , d_{ij}, data$$

$$f\left(\boldsymbol{\beta}^n\middle|\boldsymbol{\beta}, c_i\ , v_j\ , d_{ij}, data\right)$$

$$\propto \left|\Sigma_{\boldsymbol{\beta 0}}\right|^{-\frac{1}{2}} exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^n - \overline{\boldsymbol{\beta_0}}\right)' \Sigma_{\boldsymbol{\beta 0}}^{-1}\left(\boldsymbol{\beta}^n - \overline{\boldsymbol{\beta_0}}\right)\right] L(Y)$$

Here $\boldsymbol{\beta} = [\boldsymbol{\beta^c}, \boldsymbol{\beta^v}, \boldsymbol{\beta^d}]$, $\overline{\boldsymbol{\beta_0}}$ and $\Sigma_{\boldsymbol{\beta 0}}$ are diffused priors, where we set $\overline{\boldsymbol{\beta_0}}$ to be a vector of zeros and $\Sigma_{\boldsymbol{\beta 0}} = 30I$. Metropolis-Hasting algorithm is employed to randomly draw $\boldsymbol{\beta}^d$ for the new iteration from the conditional distribution described in equation above. The probability of accepting the newly drawn vector $\boldsymbol{\beta}^d$ is calculated as:

$$\Pr(\text{accept new value}) = \min\left\{\frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^d - \overline{\boldsymbol{\beta_0}}\right)' \Sigma_{\boldsymbol{\beta 0}}^{-1}\left(\boldsymbol{\beta}^d - \overline{\boldsymbol{\beta_0}}\right)\right] L(Y|\boldsymbol{\beta}^d)}{\text{xp}\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta_0}}\right)' \Sigma_{\boldsymbol{\beta 0}}^{-1}\left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta_0}}\right)\right] L(Y|\boldsymbol{\beta})}, 1\right\}$$

If the newly drawn vector is accepted, we then assign $\boldsymbol{\beta}^n = \boldsymbol{\beta}^d$.

Step 2: Generating c_i^n, v_j^n, unobserved terms for client and vendor firm, respectively.

$$f\left(c_i^n\middle|\boldsymbol{\beta}^n, c_i\ , v_j\ , d_{ij}, data\right)$$

$$\propto {\sigma_c}^{-1} exp\left[-\frac{1}{2}(c_i^n)^2 {\sigma_c}^{-1}\right] L(Y)$$

$$f\left(v_j^n\middle|\boldsymbol{\beta}^n, c_i\ , v_j\ , d_{ij}, data\right)$$

$$\propto {\sigma_v}^{-1} exp\left[-\frac{1}{2}\left(v_j^n\right)^2 {\sigma_v}^{-1}\right] L(Y)$$

Here, the two $f(\cdot \mid \cdot)$ functions are posterior distributions. Intuitively, posterior distribution can be considered to summarize your existing belief about certain distribution (prior distribution) and additional empirical data you observed (data). As a closed form solution does not exist for the two equations above, we again use Metropolis-Hasting algorithm to randomly draw from the conditional distribution specified above. We employ the method in Atchade (2006) to adaptively change the length of the steps in each iteration to help reduce the autocorrelation across MCMC iterations. The probability of accepting the newly drawn value for c_i^n, v_j^n are:

$$\Pr(\text{accept new value}) = \min\{\frac{\exp\left[-\frac{1}{2}(c_i^n)^2 {\sigma_c}^{-1}\right] L(Y|c_i^n)}{\text{eexp}\left[-\frac{1}{2}\left(c_i\ \right)^2 {\sigma_c}^{-1}\right] L(Y|c_i\ )}, 1\}$$

$$\Pr(\text{accept new value}) = \min\{\frac{\exp\left[-\frac{1}{2}\left(v_j^n\right)^2 {\sigma_v}^{-1}\right] L(Y|v_j^n)}{\text{eexp}\left[-\frac{1}{2}\left(v_j\ \right)^2 {\sigma_v}^{-1}\right] L(Y|v_j\ )}, 1\}$$

Step 3: Generating σ_c^n and σ_v^n, the standard deviation of client and vendor unobserved terms

The newly updated standard deviations for client and vendor specific unobserved terms are drawn from the distribution below:

$$\sigma_c^n \mid c_i^n \sim IW(7+N, 1+\sum_i (c_i^n)^2)$$

$$\sigma_v^n \mid v_j^n \sim IW(7+N, 1+\sum_i (v_j^n)^2)$$

Here we choose $7+N$, $1+\sum_i (c_i^n)^2$ and $1+\sum_i (v_j^n)^2$ as hyperparameters of these two distributions as they generate better predictive performance compared with other potential values. Note that hyperparameters of a MCMC are parameters of the underlying distribution from which our model parameters are generated (i.e. σ_c and σ_v). IW denotes an inverse-Wishart distribution. The reason why we chose inverse-Wishart distribution is because inverse-Wishart is the conjugate prior for the variance of a normal distribution (i.e. σ_c and σ_v). This allows us to easily simulate

values of variance (i.e. σ_c and σ_v) from a distribution with closed form, instead of using Metropolis-Hasting algorithm in Step 2.

Step 4: Generating d_{ij}^n, unobserved dyad term

$$f\left(d_{ij}^n|\boldsymbol{\beta}^n, c_i^n, v_j^n, d_{ij}, data\right)$$

$$\propto {\sigma_d}^{-1} exp\left[-\frac{1}{2}\left(d_{ij}^n\right)^2 {\sigma_d}^{-1}\right] L(Y)$$

Similar to client and vendor specific unobserved terms, we also use Metropolis-Hasting algorithm to draw from the conditional distribution above. Method proposed by Atchade (2006) is used to adaptively change the length of step in each iteration. The probability of accepting new value is:

$$\Pr(\text{accept new value}) = \min\left\{\frac{\exp\left[-\frac{1}{2}\left(d_{ij}^n\right)^2 {\sigma_d}^{-1}\right] L(Y|d_{ij}^n)}{\text{eexp}\left[-\frac{1}{2}\left(d_{ij}\right)^2 {\sigma_d}^{-1}\right] L(Y|d_{ij})}, 1\right\}$$

Step 5: Generating σ_d^n, the standard deviation of dyad unobserved term

Similar to Step 3, σ_d^n can be drawn from the following distribution:

$$\sigma_d^n |\, d_{ij}^n \sim IW(1 + N(N-1), 1 + \sum_i \sum_j (d_{ij}^n)^2)$$

Here we choose $1 + N(N-1)$ and $1 + \sum_i \sum_j (d_{ij}^n)^2$ as hyperparameters of this distribution as they generate better predictive performance compared with other potential values. IW here also denotes an inverse-Wishart distribution.

Step 6: Go back to Step 1 if the estimation is not converged.

APPENDIX II Correlation Matrix

Table A1. Correlation Matrix

	$VDegIn_{jt}$	$VDegOth_{jt}$	$VDivInd_{jt}$	$VDivServ_{jt}$	$VTAlign_{jt}$	$CDeg_{it}$	$PInter_{ijt}$	$IndTie_{ijt}$	$VCAlign_{ijt}$
$VDegIn_{jt}$	1.000	0.378	0.223	0.203	0.012	0.073	0.364	0.234	0.158
$VDegOth_{jt}$	0.378	1.000	0.358	0.336	0.010	0.008	0.259	0.371	0.090
$VDivInd_{jt}$	0.223	0.358	1.000	0.386	0.025	0.011	0.117	0.297	0.052
$VDivServ_{jt}$	0.203	0.336	0.386	1.000	0.001	0.005	0.115	0.274	0.033
$VTAlign_{jt}$	0.012	0.010	0.025	0.001	1.000	0.014	-0.006	-0.003	0.014
$CDeg_{it}$	0.073	0.008	0.011	0.005	0.014	1.000	0.098	0.302	0.025
$PInter_{ijt}$	0.364	0.259	0.117	0.115	0.006	0.098	1.000	0.361	0.189
$IndTie_{ijt}$	0.234	0.371	0.297	0.274	-0.003	0.302	0.361	1.000	0.162

$VCAlign_{ijt}$	0.158	0.090	0.052	0.033	0.014	0.025	0.189	0.162	1.000

APPENDIX III. Robustness Check with Client Firm Characteristics

In our main model, we do not control for client and vendor firm characteristics as many firms do not have historical firm-level data. In this robustness check, we include employee count and gross profits for clients who are publicly listed in North America. We present the estimation results in tables below. As we can see, the results do not change significantly.

Table A2. Robustness Check

	Coefficient
Vendor-specific Covariates	
Vendor Deg. Centrality at t-1	0.014***
Vendor Diversity (Industry) at t-1	0.049**
Vendor Diversity (Service Type) at t-1	0.026*
Vendor Task Alignment	0.095**
Dyadic-specific Covariates	
Vendor-Client Alignment	0.185***
Prior Interaction	2.281**
Prior Interaction*Vendor Degree	-0.006
Prior Interaction*Client Degree	-0.020*
Indirect Tie	0.097***
Indirect Tie *Vendor Degree	0.000
Indirect Tie *Client Degree	-0.001***
Client-specific and Control Covariates	
Client Degree at t-1	0.002
Client Profit	0.016**
Client Employee Number	0.036***
Pre 99	-0.176***
2-digit SIC codes	Included
ITO	0.211***
BPO	0.108***
#Observations	10,117

***, ** and * denote that the 99% credible interval, the 95% credible interval, and the 90% credible interval, respectively, does not include zero. The benchmark service type case is *Other*. 2-digit industry control variables have not been reported in the interest of space.

APPENDIX IV. An Example of Contract Description in Our Dataset

Company A, a provider of Internet financial technologies and solutions, has awarded a 5-year technology support contract to Company B. Company A's technologies and solutions enable financial institutions to offer online financial services to their customers. Both companies expect this to be the start of an increasingly important business relationship in line with the growth of Internet banking in Europe. IDC estimates that this contract has a 5-year life and a value of $20-25 million. Contract Responsibilities: Through its relationship with Company A, Company B will provide desktop and server infrastructure support to Company C, to which Company A already provides Internet financial technologies, solutions and Web hosting facilities. Company B will also work with Company A to assist Company C as it plans and implements its future technology strategy including the management of other third party information technology service providers. Company B will provide services and operate in Dublin. The services will include support operations onsite in Ireland, including servicing-the UK, Germany and Singapore. This contract represents a key technology partnership for both Company B and Company A.